Please cite this not peer-reviewed draft/preprint as: Rzymski, Christoph, Tiago Tresoldi, Simon Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, Johann-Mattis List (2019): The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. Max Planck Institute for the Science of Human History: Jena.

The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies

*Christoph Rzymski[1], *Tiago Tresoldi[1], Simon Greenhill[1,2], Mei-Shin Wu[1], Nathanael E. Schweikhard[1], Maria Koptjevskaja-Tamm[3], Volker Gast[4], Timotheus A. Bodt[5], Abbie Hantgan[6], Gereon A. Kaiping[7], Sophie Chang[8], Yunfan Lai[1], Natalia Morozova[1], Heini Arjava[9], Nataliia Hübler[13], Ezequiel Koile[1], Steve Pepper[10], Mariann Proos[11], Briana Van Epps[12], Ingrid Blanco[4], Carolin Hundt[4], Sergei Monakhov[4], Kristina Pianykh[4], Sallona Ramesh[4], Russell D. Gray[1], *Robert Forkel[1], *Johann-Mattis List[1].

July 25, 2019

Abstract

Advances in computer-assisted linguistic research are greatly influencing and reshaping linguistic investigation. With the increasing availability of interconnected datasets created and curated by researchers, more and more interwoven questions can now be investigated. Such advances, however, are bringing high requirements in terms of rigorousness for preparing and curating datasets. In this work we present CLICS, a Database of Cross-Linguistic Colexifications which aims to both tackle interdisciplinary and interconnected research questions as well as showcasing best practices in preparing data for cross-linguistic research. This is done by addressing shortcomings of an earlier version of the database, CLICS^2, and supplying an updated version with CLICS^3 which massively increases the size and scope of the project. We provide tools and guidelines for this purpose and discuss insights resulting from organizing student tasks for database updates.

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, [2]ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra, [3]Stockholm University, [4]Friedrich Schiller University, Jena, [5]SOAS, London, [6]CNRS LLACAN, [7]University of Leiden, [8]Independent English-Chinese Translator and linguistic researcher, [9]University of Helsinki, [10]University of Oslo, [11]University of Tartu, [12]Lund University, *Corresponding author: {rzymski,tresoldi,forkel,list}@shh.mpg.de.

1 Background and Introduction

The quantitative turn in historical linguistics and linguistic typology has drastically changed the way in which scholars create, use, and share linguistic information. Along with the constantly growing amount of digitally available data for the world's languages, we find a drastic increase in the application of new quantitative techniques. While most of the new methods are inspired by neighboring disciplines and general-purpose frameworks, such as evolutionary biology [1, 2], machine learning [3, 4], or statistical modeling [5, 6], the particularities of cross-linguistic data often necessitate a specific treatment of materials (reflected in recent standardization efforts [7, 8]) and methods (illustrated by the development of new algorithms tackling specifically linguistic problems [9, 10]).

This increase in quantitativeness becomes particularly clear in studies on *cross-linguistic semantics* (or *semantic typology*), which investigate how languages distribute meanings across their vocabularies. Although questions concerning such categorizations across human languages have a long-standing tradition in linguistics and philosophy [11, 12], global-scale studies have long been restricted to certain recurrent semantic fields, such as *color terms* [13, 14], *kinship terms* [15, 16], and *numeral systems* [17], invariably involving smaller amounts of data with lower coverage of linguistic diversity in terms of families and geographic areas.

Along with improved techniques in data creation and curation, advanced computational methods have opened new possibilities for research in this area. One example is the *Database of Cross-Linguistic Colexifications* (CLICS, `https://matthew.clld.org/clics`), first published in 2014 [18], which offers a framework for the computer-assisted collection, computation, and exploration of worldwide patterns of cross-linguistic "colexifications". The term *colexification* [19] refers to instances where the same word expresses two or more comparable concepts [20, 79], such as in the common case of *wood* and *tree* "colexifying" in languages like Russian (both expressed by the word "дерево" [derevo]) or Nahuatl ("k^{w}owi-t"). By harvesting colexifications across multiple languages, with recurring patterns potentially reflecting universal aspects of human perception and cognition, researchers are able to identify cross-linguistic polysemies without resorting to intuitive decisions about the motivation for such identities.

The CLICS project reflects the rigorous and transparent approaches to standardization and aggregation of linguistic data, allowing to investigate colexifications by means of semantic networks involving global occurrences, as in the example of Figure 1, mostly by reusing data originally collected for historical linguistics. Its framework is designed, along with the corresponding interfaces, to facilitate the exploration and testing of alleged cross-linguistic polysemies [21] and areal patterns [22]. The project is rapidly becoming a popular tool not only for examining cross-linguistic patterns, particularly those involving unrelated languages, but also for conducting new research in fields not strictly related to semantic typology [23–27].

A second version of the CLICS database was published in 2018, revising and drastically increasing the amount of cross-linguistic data [28]. Such improve-

ments were made possible by an enhanced strategy of *data aggregation* relying on the standardization efforts of the Cross-Linguistic Data Formats initiative [7] (CLDF, https://cldf.clld.org), which provides standards, tools, and best practice examples for the promotion of linguistic data which is FAIR in the sense of Wilkinson [29]: *findable*, *accessible*, *interoperable*, and *reusable*. By adopting these principles and coding independently published cross-linguistic datasets according to the specifications recommended by the CLDF initiative, it was possible to increase the amount of languages from less than 300 to more than 1000, while expanding the number of concepts in the study from 1200 to more than 1500.

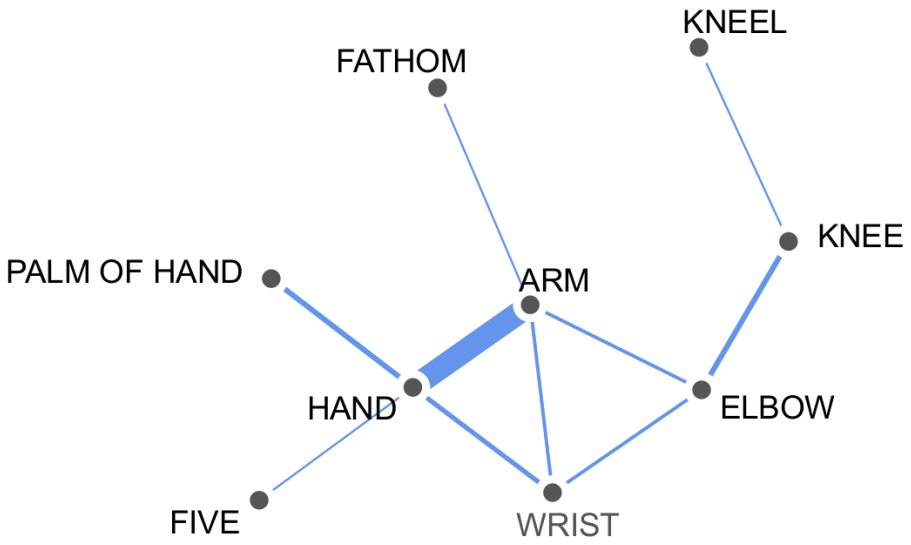

Figure 1: Example of a colexification network [30]. A strong link between ARM and HAND is shown, indicating that in many languages both concepts are expressed with the same word; among others, weaker links between concepts HAND and FIVE, explainable by the number of fingers in a hand, and ELBOW and KNEE, explainable as both being joints, can also be observed.

A specific shortcoming of this second release of CLICS was that, despite being based on CLDF format specifications, it did not specify how data conforming to such standards could be created in the first place. Thus, while the CLDF data underlying CLICS² are findable, accessible, interoperable, and reusable, the procedures involving their creation and expansion were not necessarily easy or evident.

In order to tackle this problem, we have developed guidelines and software tools which help greatly with the conversion of existing linguistic dataset into the CLDF format. We tested the suitability of our new curation framework by conducting two student tasks in which students with a background

in linguistics helped us to convert and integrate data from different sources into our database. We illustrate the efficiency of this workflow by providing an updated version of our data, which increases the number of languages from 1220 to 2955 and the number of concepts from 2487 to 2811. In addition, we also increased and enhanced the transparency, flexibility, and reproducibility of the workflow by which CLDF datasets are analyzed and published within the CLICS framework, publishing a testable virtualized container [31] that can be freely used on-line in the form of a *Code Ocean capsule* (https://codeocean.com/capsule/4564348).

2 Methods

2.1 Create and curate data in CLDF

The CLDF initiative promotes principles, tools, and workflows to make data cross-linguistically compatible and comparable, facilitating interoperability without strictly enforcing it or requiring linguists to abandon their long-standing data management conventions and expectations. Key aspects of the data format advanced by the initiative are an exhaustive and principled use of reference catalogs, such as Glottolog [80] for languages and Concepticon [32] for comparative concepts, along with standardization efforts like the Cross-Linguistic Transcription Systems (CLTS) for normalizing phonological transcriptions [8].

Preparing data for CLICS starts with obtaining and expanding raw data, often in the form of Excel tables (or similar table formats) as shown in Figure 2.

	A	B	C	D	E	F
1	ID	Parameter	English	Chinese	Pinyin	Pla
2	0	749	fly	飞	fēi	bjy¹
3	1	813	straight	直 (棍子很直)	zhí	dʒaŋ²
4	2	403	bracelet	手镯	shǒuzhuó	lɐ̯ˡgoŋ¹
5	3	67	afternoon	下午	xiàwǔ	a¹mo²kʰi̯ˡ
6	4	68	dusk/evening	黄昏	huánghūn	a¹mo²ɣi¹
7	5	68.1	dusk/evening	黄昏	huánghūn	a¹mo²kʰi̯ˡ
8	6	737	roll	滚 (石头滚)	gǔn	ʔly³/²
9	7	234	mouth	嘴	zuǐ	kʰa²pɐ̯ˡ
10	8	234.1	mouth	嘴	zuǐ	kʰa²pɐ̯ˡ
11	9	235	lips	嘴唇	zuǐchún	mi²la³
12	10	235.1	lips	嘴唇	zuǐchún	mi²be³dzi³
13	11	851	cool	凉快	liángkuài	
14	12	415	shoulder bag	肩袋	jiāndài	ta²la²py³
15	13	191.2	grass	草	cǎo	mɐ̯ˡ

Figure 2: Raw data as a starting point for applying our workflow (a snippet from the `yanglalo` dataset [81]).

By using our sets of tools, data can be enriched, cleaned, improved, and made ready for usage in multiple different applications, both current, such as CLICS, or future ones using compliant data.

This toolbox of components supports the creation and release of CLDF datasets through a fully integrated workflow comprising six fundamental steps (as illustrated in Figure 3). First, (1) scripts prepare raw data from sources for digital processing, leading the way to the subsequent catalog cross-referencing at the core of CLDF. This task includes the steps of (2) referencing sources in the BibTeX format, (3) linking languages to Glottolog, and (4) mapping concepts to Concepticon. To guarantee straightforward processing of lexical entries by CLICS and other systems, the workflow might also include a step for (5) cleaning lexical entries of systematic errors and artifacts from data conversion. Once the data have been curated and the scripts for workflow reproducibility are completed, the dataset is ready for (6) public release as a package relying on the `pylexibank` library, a step that includes publishing the CLDF data on Zenodo and obtaining a DOI.

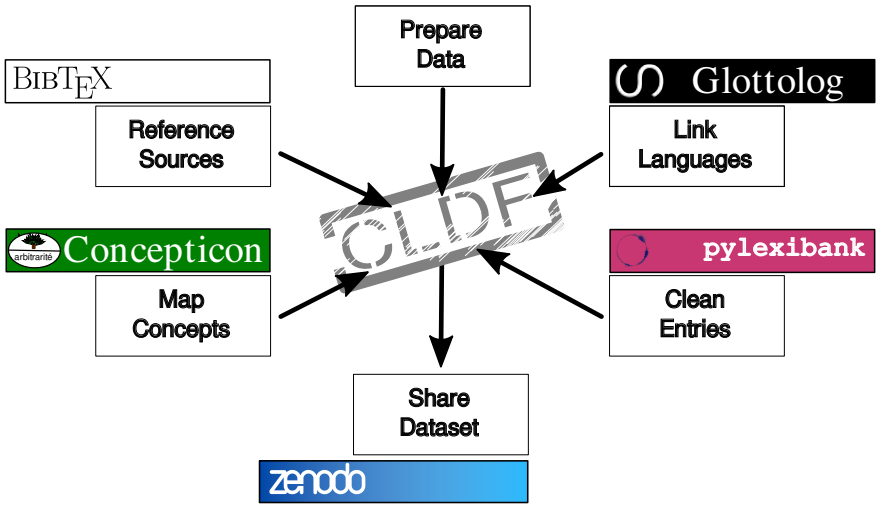

Figure 3: A diagram representing the six fundamental steps of a CLDF dataset preparation workflow.

The first step in this workflow, preparing source data for digital processing, varies according to the characteristics of each dataset. The procedure ranges from the digitization of data collections only available as book scans or even fieldwork notes (using software for optical character recognition or manual labor, as done for the `beidasinitic` dataset [82]), via the re-arrangement of data distributed in word processing or spreadsheet formats such as `docx` and `xlsx` (as in the `castrosui` dataset [83]), up to the extraction of data from websites (as done for `diacl` [84]). In quite a few cases, scholars helped us by sharing fieldwork data (`yanglalo` [85], `bodtkhobwa` [86]), or providing the unpublished data underlying a previous publication (e.g. `satterthwaitephillipstb` [87]). In other cases, we profited from the digitization efforts of large documentation

projects such as STEDT [33] (the source of the `suntb` [88] dataset via [34]) and Northeuralex [89].

In the second step, we identify all relevant sources used to create a specific dataset and store them in BibTeX format, the standard for bibliographic entries required by CLDF. This is done on a per-entry level, guaranteeing that for each data point it will always be possible to identify the original source; the `pylexibank` library will dutifully list all rows missing bibliographic references, treating them as incomplete entries. Given the large amount of bibliographic entries on language resources provided by aggregators like Glottolog [80], this step is usually straightforward, although it may require more effort when the original dataset does not properly reference its sources.

The third and fourth steps comprise linking language varieties and concepts used in a dataset to the Glottolog and the Concepticon catalogs, respectively. Both such references are curated on publicly accessible GitHub repositories, allowing researchers to submit or request changes and obtain a local copy of the entire catalogues. In both cases, on-line interfaces are available for open consultation (at `https://glottolog.org/` and `https://concepticon.clld.org/`, respectively). While these linking tasks require some linguistic expertise, such as for the distinction of the language varieties involved in a study, both projects provide libraries and tools for semi-automatic mapping that facilitate and speed up the tasks. For example, the mapping of concepts was tedious in the past when the entries in the published concept lists differed too much from proper glosses, such as when part-of-speech information was included along with the actual meaning or translation, often requiring a meticulous comparison between the published work and the corresponding concept lists. However, the second version of Concepticon introduced new methods for semi-automatic concept mapping through the `pyconcepticon` package, which can be invoked from the command-line, besides a lookup-tool allowing to quickly search concepts by fuzzy matching of elicitation glosses. Depending on the size of a concept list, this step can still take several hours, but the lookup procedure has been drastically improved in the last version, also due to the steadily increasing number of concepts and concept lists being added.

In a fifth step, we use the `pylexibank` API to clean lexical entries from systematic errors. This API allows users to convert data in raw format – when bibliographic references, links to Glottolog, and mappings to Concepticon are provided – to proper CLDF datasets. Given that linguistic datasets are often inconsistent regarding lexical form rendering, the programming interface is used to automatically clean the entries by (a) splitting multiple synonyms from their original *value* into unique *forms* each, (b) deleting brackets, comments, and other parts of the entry which do not reflect the original word form, but rather authors' and compilers' comments, (c) making a list of entries to ignore or manually correct, in case the automatic routine does not capture all idiosyncrasies, and (d) using explicit mapping procedures for converting from orthographies to phonological transcriptions. The resulting CLDF dataset contains both the original and unchanged textual information, labeled *Value*, and its processed version, labeled *Form*, explicitly informing what is taken from the

original source and what results from our manipulations, always allowing to compare the original and curated state of the data. Even when the original is clearly erroneous, for example due to misspellings, the *Value* is left unchanged and the information is corrected only in the *Form*.

As a final step, CLDF datasets are publicly released. The datasets live as individual GitHub repositories that can be anonymously cloned. A dataset package contains all the code and data resources required to recreate the CLDF data locally, as well as interfaces for easily installing and accessing the data in any Python environment. Packages can be frozen and subsequently released on platforms like Zenodo, supplying them with persistent identifiers and archiving for reuse and data provenance. The datasets for CLICS[3], for example, are aggregated within the Zenodo community at `https://zenodo.org/communities/clics/`.

Besides the transparency in line with the best practices for open access and reproducible research, the efficiency of this workflow and of the underlying initiative can be demonstrated by the improvements in the CLICS project. The first version [18] was based on only four datasets publicly available at the time of its development. The project was well received and reviewed, particularly due to the release of its aggregated data in an open and reusable format, but as a cross-linguistic project it suffered from several shortcomings in terms of data *coverage*, being heavily biased towards European and South-East Asian languages. The second version of CLICS [30] combined 15 different datasets already in CLDF format, making data reuse much easier, while also considerably increasing quality and coverage of the data. The new version, as detailed in Table 2, doubles the number of datasets without particular needs for changes in CLICS itself. The project is fully integrated with LEXIBANK and with the CLDF libraries, and, as a result, when a new dataset is published, it can be installed to any local CLICS setup which, if instructed to rebuild its database, will incorporate the new information in all future analyses. Likewise, it is easy to restrict experiments by loading only a selected subset of the installed datasets. The rationale behind this workflow is shared by similar projects in related fields (e.g. computational linguistics), where data and code are to be strictly separated, allowing research to test different approaches and experimental setups without much effort.

2.2 Colexification analysis with CLICS

CLICS is distributed as a standard Python package comprising the `pyclics` programming library and the `clics` command-line utility. Both the library and the utility require a CLICS-specific lexical database; the recommended way of creating one is through the `load` function: calling `clics load` from the command-line prompt will create a local SQLITE database for the package and populate it with data from the installed Lexibank datasets. While this allows researchers with specific needs to select and manually install the datasets they intend, for most use cases we recommend using the curated list of datasets distributed along with the project and found in the `clicsthree/datasets.txt`

file. The list follows the structure of standard `requirements.txt` files and the entire set can easily be installed with the standard `pip` utility.

The installation of the CLICS tools is the first step in the workflow for conducting colexification analyses. The following points describe the additional steps, and the entire workflow is illustrated in the diagram of Figure 4.

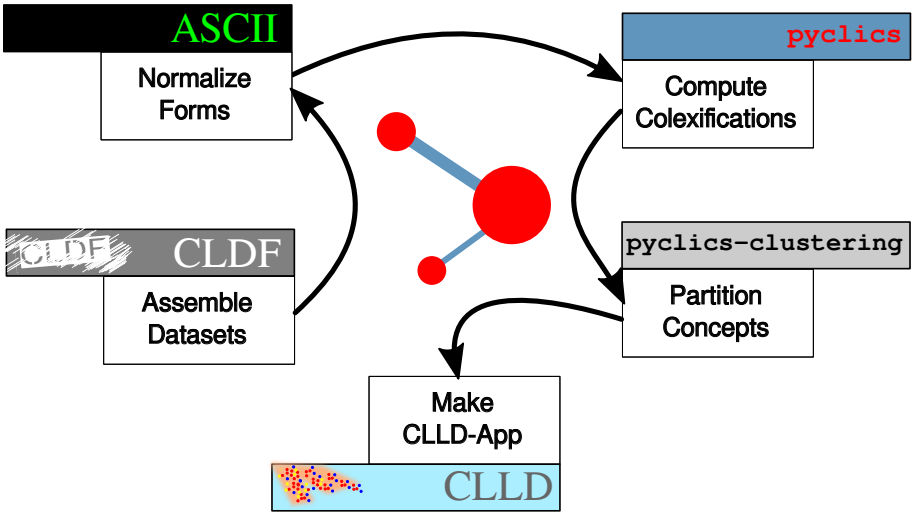

Figure 4: A diagram representing the workflow for installing, preparing, and using CLICS.

Firstly, a set of CLDF datasets is assembled into a CLICS database. Once the database has been generated, a colexification graph can be computed. As already described when introducing CLICS [18] and CLICS² [28], a colexification graph is an undirected graph in which nodes represent comparable concepts and edges express the colexification weight between the concepts they link: for example, *wood* and *tree*, two concepts that as already mentioned colexify in many languages, will have a high edge weight, while *water* and *dog*, two concepts without a single instance of lexical identity in our data, will have an edge weight of zero.

Secondly, all forms in the database are normalized. Normalized forms are forms reduced to more basic and comparable versions by additional operations of string processing, removing information such as morpheme boundaries or diacritics, eventually converting the forms from their Unicode characters to the closest ASCII approximation by means of the `unidecode` library [35].

Thirdly, colexifications are then computed by taking the combination of all comparable concepts found in the data and, for each language variety, comparing for equality the cleaned forms that express both concepts (the comparison might involve more than two words, as it is common for sources to list synonyms). Information on the colexification for each concept pair is collected both in terms of languages and language families, given that patterns found

across different language families are more likely to be a polysemy stemming from human cognition than patterns due to vertical transmission or random resemblance. Cases of horizontal transmission ("borrowings") might confound the clustering algorithms to be applied in the next stage, but our experience has shown that colexifications are actually a useful tool for identifying candidates of horizontal transmission and areal features. Once the number of matches has been collected, edge weights are adjusted according to user-specified parameters, for which sensible defaults are provided.

The output of running CLICS[3] with default parameters, reporting the most common colexifications and their counts for the number of language families, languages, and words, is shown in Table 1.

Concept A	Concept B	Families	Languages	Words
MOON	MONTH	57	321	329
TREE	WOOD	57	305	412
FINGERNAIL	CLAW	55	222	230
KNIFE (FOR EATING)	KNIFE	51	268	285
SON-IN-LAW (OF MAN)	SON-IN-LAW (OF WOMAN)	49	261	284
LANGUAGE	WORD	49	117	122
LEG	FOOT	48	296	313
SKIN	BARK	48	200	218
LISTEN	HEAR	48	114	117
DAUGHTER-IN-LAW (OF MAN)	DAUGHTER-IN-LAW (OF WOMAN)	47	234	261

Table 1: The ten most common colexifications for CLICS[3], as the output of command `clics colexifications`.

Finally, the graph data generated by the colexification computation, along with the statistics on the score of each colexification and the number of families, languages, and words involved, can be used in different quantitative analyzes, e.g. clustering algorithms to partition the graph in "subgraphs" or "communities". A sample output created with infomap clustering and a family threshold of 3 is illustrated in Figure 5.

Our experience with CLICS confirms that, as in most real-world networks and particularly in social ones, nodes from colexification studies are not evenly distributed, but concentrate in groups of relatively high density that can be identified by the most adopted methods [36, 37] and even by manual inspection: while some nodes might be part of two or more communities, the clusters detected by the clustering of colexification networks are usually quite distinct one from the other [38, 39]. These can be called "semantic communities", as they tend to be clearly linked in terms of semantic proximity, establishing relationships that, in most cases, linguists have described as acceptable or even expected, with one or more central nodes acting as "centers of gravity" for the cluster: one example is the network already shown in Figure 1, oriented towards the anatomy of human limbs and centered on the strong *arm-hand* colexification.

CLICS tools provide different clustering methods (see 5) that allow to identify clusters for automatic or manual exploration, especially when using its graphical interface. Both methods not only identify the semantic communities but also collect complementary information allowing to appropriately name each one after the semantic centers of the subgraph.

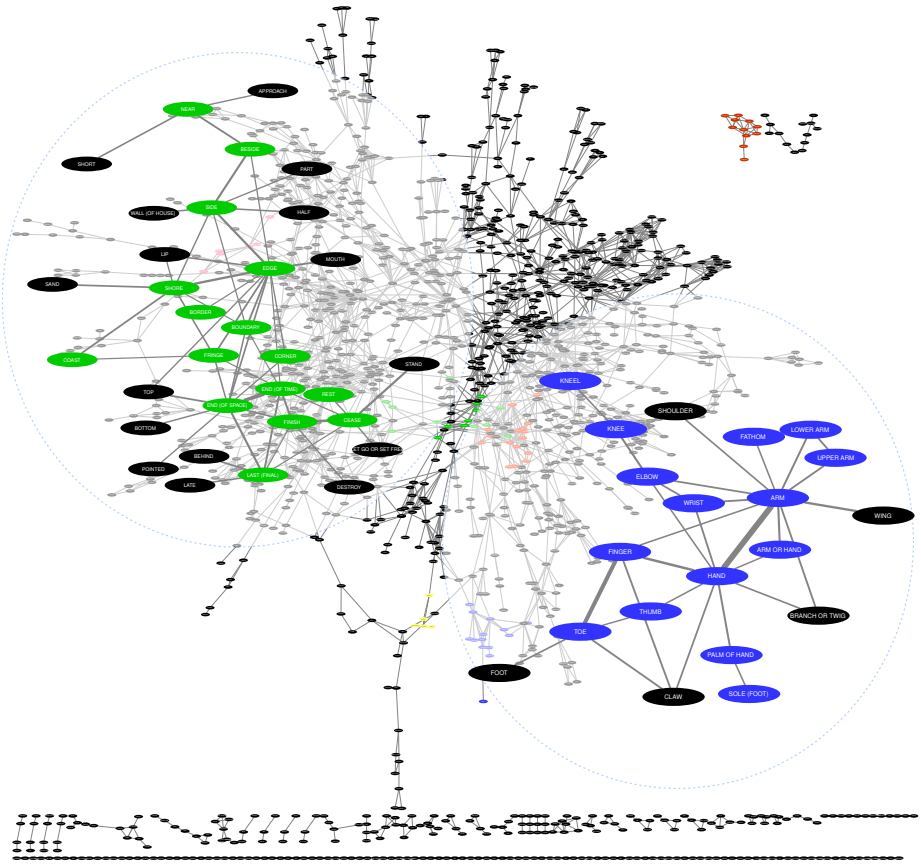

Figure 5: Colexification clusters in CLICS³.

The command-line utility can perform clustering through its `cluster` command followed by the name of the algorithm to use (a list of the algorithms is provided by the `clics cluster list` command). For example, `clics cluster infomap` will cluster the graph with the *infomap* algorithm [40], in which community structure is detected by means of random walks (with a community mathematically defined as a group of nodes with more internal than external connecting edges). After clustering, additional summary statistics can be obtained from the `clics graph-stats` command: for standard CLICS³ with default parameters and clustering with the recommended and default *infomap* algorithm, the process results in 1624 nodes, 2871 edges, 96 components, and 256 communities.

The data generated by following the workflow outlined in 4 can be used in multiple different ways (see 5), e.g. for preparing a web-based representation of the computed data using the CLLD [41] toolkit.

3 Data records

CLICS³ is distributed with 30 different datasets, as detailed in Table 2, of which half were included for this new release. Most datasets were originally collected for purposes of language documentation and historical linguistics, such as `bodtkhobwa` [42], while a few were generated from existing lexical collections, as `logos` [43], or from previous linguistic studies, as in the case of `wold` [44]. Datasets were selected for inclusion either due to interest for historical linguistics, to the desire of expanding the coverage of CLICS² in terms of linguistic families and areas, or due to on-going collaborations with the authors of the studies.

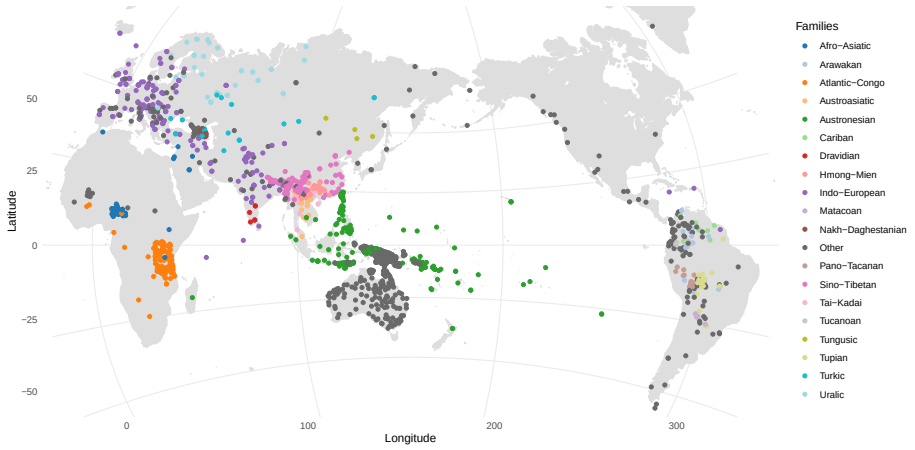

Figure 6: Distribution of language variaties in CLICS³.

4 Technical validation

In order to investigate to which degree our enhanced workflows would improve the efficiency of data creation and curation within the CLDF framework, we conducted two tests. First, we tested the workflow ourselves by actively searching for new datasets which could be quickly added to our framework, noting improvements that could be made for third-party usage and public release. Second, once such experience was maturing, two student tasks with the goal of adding new datasets to the CLICS database were organized, both involving the delegation of parts of the workflow to students of Linguistics. In the following

	Dataset	Glosses	Concepticon	Varieties	Glottocodes	Families	New	References
1	abrahammonpa	304	304	26	15	2	Yes	[45] [90]
2	allenbai	497	496	9	9	1		[46] [91]
3	bantubvd	420	415	10	10	1		[47] [92]
4	beidasinitic	713	713	18	18	1		[48] [82]
5	bodtkhobwa	543	529	8	8	1	Yes	[42] [86]
6	bowernpny	338	338	170	168	1		[49] [93]
7	castrosui	507	498	16	3	1	Yes	[50] [83]
8	chenhmongmien	783	737	22	20	1	Yes	[51] [94]
9	diacl	537	537	368	348	25	Yes	[52] [84]
10	halenepal	720	678	13	13	2	Yes	[53] [95]
11	hantganbangime	299	299	22	22	5	Yes	[54] [96]
12	hubercolumbian	348	344	69	65	16		[55] [97]
13	ids	1310	1306	320	275	60		[56] [98]
14	kraftchadic	428	428	67	60	3		[57] [99]
15	lexirumah	604	602	179	140	4	Yes	[58] [100]
16	logos	707	707	5	5	1	Yes	[43] [101]
17	marrisonnaga	667	645	36	35	1	Yes	[59] [102]
18	mitterhoferbena	342	335	13	13	1	Yes	[60] [103]
19	naganorgyalrongic	960	870	10	8	1	Yes	[61] [104]
20	northeuralex	950	949	107	107	21		[62] [89]
21	robinsonap	392	392	13	13	1		[63] [105]
22	satterthwaitetb	418	418	18	18	1		[64] [87]
23	sohartmannchin	279	279	6	5	1	Yes	[65] [106]
24	suntb	915	904	49	49	1		[34] [88]
25	tls	1101	808	120	97	1		[66] [107]
26	transnewguineaorg	904	865	1004	760	106	Yes	[67] [108]
27	tryonsolomon	317	314	111	96	5		[68] [109]
28	wold	1460	1458	41	41	24		[44] [110]
29	yanglalo	884	851	7	7	1	Yes	[81] [85]
30	zgraggenmadang	309	306	98	98	1		[69] [111]
TOTAL			2809	2893	2135	200		

Table 2: List of datasets included in CLICS³, along with individual counts for glosses ("Glosses"), concepts mapped to Concepticon ("Concepts"), language varieties ("Varieties"), language varieties mapped to Glottolog ("Glottocodes"), and language families ("Families"). New datasets included for the CLICS³ release are also indicated; references first list the original source for the data, and second the corresponding CLDF dataset. The references next to the dataset names refer to original source; information on the dataset compilers is provided in the Zenodo repository along with the bibliographic entry in the references.

paragraphs, we will quickly discuss our experiences with such tasks, besides presenting some detailed information on the notable differences between CLICS² and the improved CLICS³ resulting from both tests.

4.1 Workflow validation

In order to validate the claims of improved reproducibility and the general validity of the workflow for preparing, adding, and analyzing new datasets, we conducted two student tasks in which participants at Ph. D. and undergraduate level were asked to contribute to CLICS³ by using the tools we developed. The first student task was carried out as part of a seminar for doctoral students on *Semantics in Contact*, taught by M. Koptjevskaja-Tamm (MKT) as part of a summer school of the Societas Linguistica Europaea (August 2018, University of Tartu). The second task was carried out as part of an M.A. level course on *Methods in Linguistic Typology*, taught by V. Gast (VG) as a regular seminar at the Friedrich Schiller University (Jena) in the winter semester of 2018/2019.

MKT's group was first introduced to CLICS², to the website accompanying the CLICS project, and to the general ideas behind a colexification database. This helped shaping a better understanding of what is curated in the context of CLICS. In a second step, we provided a task description tailored for the students, which was presented by MKT. In a shortened format, it consisted of (1) general requirements for CLICS datasets (as described in previous sections), (2) steps for digitizing and preparing data tables (raw input processing), (3) Concepticon linking (aided by semi-automatic mapping), (4) Glottolog linking (identifying languages with Glottocodes), (5) providing bibliographic information with BibTeX, (6) providing provenance information as well as verbal descriptions of the data.

The students were split in five groups of two people, and each group was tasked with carrying out one of the six tasks for a specific dataset we provided. The students were not given strict deadlines, but we informed them that they would be listed as contributors to the next update of the CLICS² database if they managed to provide the data up to two months after the task was introduced to them. While the students were working on their specific tasks, we provided additional help by answering specific questions, such as regarding the detailed mapping of certain concepts to Concepticon, via email.

All of the student groups managed to finish their tasks successfully, with only minor corrections and email interactions from our side. The processed data provided by the students lead to the inclusion of five new datasets to CLICS³: `castrosui` [83], a collection of Sui dialects of the Tai-Kadai family spoken in Southern China, `halenepal` [95], a large collection of languages from Nepal, `marrisonnaga` [102], a collection of Naga languages (a branch of the Sino-Tibetan family), `mitterhoferbena` [103], a collection of Bena dialects spoken in Tanzania, and `yanglalo` [85], a dataset of regional varieties of Lalo (a Loloish language cluster spoken in Yunnan, part of the Sino-Tibetan family).

A similar approach was taken by VG and his group of students, with special emphasis being placed on the difficulties and advantages of a process for col-

laborative and distributed data preparation. They received instruction material similar to that of MKT's group, but more nuanced towards the dataset they were asked to work with, namely `diacl` [84] (`https://diacl.ht.lu.se/`), a collection of linguistic data from 26 large language families all over the world. Pre-processed data was provided by us and special attention was paid to the process of concept mapping.

In summary, the outcome of the workflow proposed was positive for both groups, and the data produced by the students and their supervisors helped us immensely with extending CLICS³. Some students pointed us to problems in our software pipeline, such as missing documentation on dependencies in our installation instructions. Difficulties during the process of concept mapping were also indicated, such as problems arising from insufficient concept definitions for linking of elicitation glosses to concept sets. We have addressed most of these problems and hope to obtain more feedback from future users in order to further enhance our workflows.

4.2 CLICS³ validation

The technical validation of CLICS³ is based on functions for deconstructing forms and consequences of this for mapping and finding colexifications. If we compare the data status of CLICS² with the amount of data available with the release of CLICS³, we can see a drastic increase in data, both with respect to the number of languages being covered by CLICS³, and with respect to the total number of concepts included now. When looking at the detailed comparisons in Figure 7, however, we can see that the additions of data occurred in different regions of the world. While we note a drastic increase of data points in Papunesia, a point of importance for better coverage of "hot spots" [70], and a moderate increase in Eurasia, the data is unchanged in Africa, North America, and Australia, and has only slightly increased in South America. As can be easily seen from Figure 6, Africa and North America are still only sparsely covered in CLICS³. Future work should try to target these regions specifically.

While this shows, beyond doubt, that our data aggregation strategy based on transparent workflows that create FAIR data was, by and large, successful, it is important to note that the *average mutual coverage*, which is defined as the average number of concepts for which two languages share a translation [28, 71], is rather low. This, however, is not surprising, given that the original datasets were collected for different purposes. While low or skewed coverage of concepts is not a problem for the original purpose of CLICS, which is still mostly used as a tool for the manual inspection of colexifications, it should be made very clear that quantitative approaches dealing with CLICS² and CLICS³ need to explicitly control for missing data.

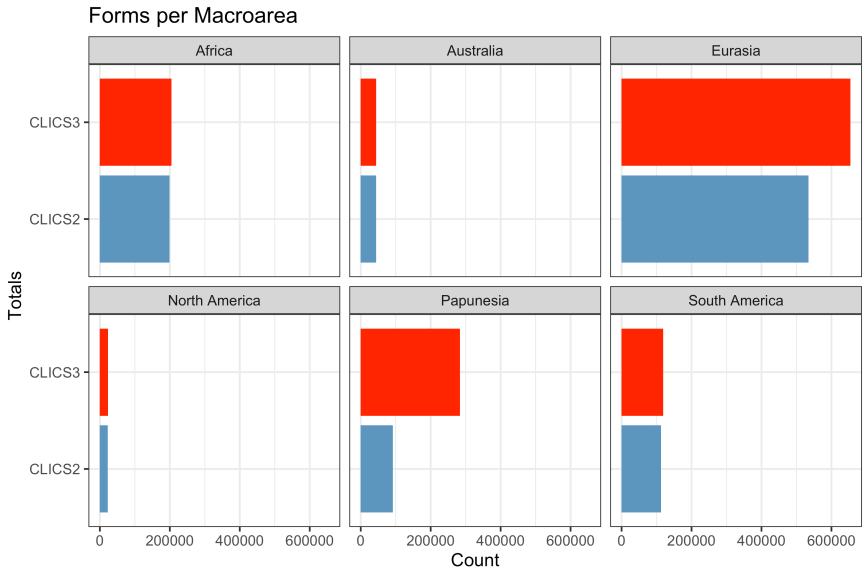

Figure 7: Increase in data points (*values*) for CLICS³.

5 Usage notes

The CLICS pipeline produces a number of artifacts that can serve as an entry point for researchers: a locally browsable interface, well-suited for exploratory research, a SQLite database containing all data points, languoids and additional information, and colexification clusters in the Graph Modelling Language (GML [72]).

The SQLite database can easily be processed with programming languages like R and Python, while the GML representation of CLICS colexification graphs is fully compatible with tools for advanced network analyses, e.g. Cytoscape [73]. Additionally, researchers have the choice between different clustering algorithms (currently supported and implemented: highly connected subgraphs [74], infomap or map equation [40], Louvain modularity [75], hierarchical clustering [76], label propagation [77], and connected component clustering [78]) and can easily plug-in and experiment with different clustering techniques using the `pyclics-clustering` package (https://github.com/clics/pyclics-clustering). A sample workflow is also illustrated in the Code Ocean capsule for this publication (https://codeocean.com/capsule/4564348). For easier accessibility, CLICS data can also be accessed on the web with our CLICS CLLD app, available on https://matthew.clld.org/clics.

Acknowledgements

This research would not have been possible without the generous support by many institutes and funding agencies. TT, MSW, NES, YL, and JML were funded by the the ERC Starting Grant 715618 Computer-Assisted Language Comparison (http://calc.digling.org). SJG was supported by the Australian Research Council's Discovery Projects funding scheme (project number DE 120101954) and the ARC Center of Excellence for the Dynamics of Language grant (CE140100041). We are also very grateful for the help and data provided by many researchers, amongst them: Cathryn Yang for [85], Andy Castro for [83], Michael Cysouw for help with the digitization of [82], [97], and [99], Claire Bowern for [93], Gerd Carling for [84], Doug Cooper for [94], and the STEDT project for providing digital versions of the data underlying [95], [102], and [104].

Author contributions

RF, JML, CR, TT, and SJG initialized the study. CR, JML, and TT wrote the first draft. SJG, RF, and RDG revised the first draft. CR, JML, MSW, NM, NES, RF, SJG, and TT provided datasets in CLDF format and helped with data curation. CR, JML, RF, SJG, and TT developed software for the CLICS pipeline. RF wrote the code for the CLLD application. AH, GK, TB, and SC contributed data. MKT, CR, and JML conducted student task 1, VG, CR, and JML conducted student task 2. BVE, EK, HA, MP, NH, and SP participated in student task 1, CH, KP, IB, SM, and SR participated in student task 2. All authors approved the final version of the manuscript.

References

Works

1. Atkinson, Q. D. & Gray, R. D. Curious Parallels and Curious Connections: Phylogenetic Thinking in Biology and Historical Linguistics. *Systematic Biology* **54,** 513–526 (2005).
2. List, J.-M., Pathmanathan, J. S., Lopez, P. & Bapteste, E. Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics. *Biology Direct* **11,** 1–17. http://biologydirect.biomedcentral.com/articles/10.1186/s13062-016-0145-2 (2016).
3. Rama, T. *Siamese convolutional networks for cognate identification* in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* COLING 2016Osaka, Dec. 11–17, 2016 (2016), 1018–1027.

4. Rama, T. & List, J.-M. *An automated framework for fast cognate detection and Bayesian phylogenetic inference in computational historical linguistics* in *57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, forthcoming), 1–11.
5. Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. & Christiansen, M. H. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Science of the United States of America* **113,** 10818–10823 (2016).
6. Bromham, L., Hua, X., Fitzpatrick, T. G. & Greenhill, S. J. Rate of language evolution is affected by population size. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 2097–2102 (2015).
7. Forkel, R. *et al.* Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* **5,** 1–10 (2018).
8. Anderson, C. *et al.* A Cross-Linguistic Database of Phonetic Transcription Systems. *Yearbook of the Poznań Linguistic Meeting* **4,** 21–53 (2018).
9. List, J.-M. *Sequence comparison in historical linguistics* (Düsseldorf University Press, Düsseldorf, 2014).
10. List, J.-M. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics* **1,** 137–161 (2019).
11. Bach Emmon; Chao, W. in *Language Universals* (eds Christiansen, M. H., Collins, C. & Edelman, S.) 1–33 (Oxford Scholarship Online, Oxford, UK, 2009).
12. Evans, N. in *The Oxford Handbook of linguistic typology* (ed Sung, J. J.) 504–533. (Oxford University Press, Oxford, 2011).
13. Berlin Brent; Kay, P. *Basic Color Terms: Their Universality and Evolution* (University of California Press, Berkeley, 1969).
14. Kay Paul; McDaniel, C. K. The linguistic significance of the meanings of basic color terms. *Language* **72,** 522–78 (1978).
15. Nerlove Sara; Romney, A. K. Sibling terminology and cross-sex behavior. *American Anthropology* **69,** 179–87 (1967).
16. Murdock, G. P. Kin term patterns and their distribution. *Ethnology* **9,** 165–208 (1970).
17. Greenberg, J. H. in *Universals of Human Language, vol. 3: Word Structure* (ed Greenberg, J. H.) 249–95 (Stanford University Press, Stanford, CA, 1978).
18. List, J.-M., Mayer, T., Terhalle, A. & Urban, M. *CLICS: Database of Cross-Linguistic Colexifications* `http://clics.lingpy.org` (Forschungszentrum Deutscher Sprachatlas, Marburg, 2014).
19. François, A. in *From polysemy to semantic change* (ed Vanhove, M.) 163–215 (Benjamins, Amsterdam, 2008).

20. Haspelmath, M. Comparative concepts and descriptive categories. *Language* **86,** 663–687 (2010).
21. Urban, M. Asymmetries in overt marking and directionality in semantic change. *Journal of Historical Linguistics* **1,** 3–47 (2011).
22. Schapper, A., Roque, L. S. & Hendery, R. in *The lexical typology of semantic shifts* (eds Juvonen, P. & Koptjevskaja-Tamm, M.) 355–422 (De Gruyter Mouton, Berlin and Boston, 2016).
23. Brochhagen, T. *Improving coordination on novel meaning through context and semantic structure* in *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning* (2015), 74–82.
24. Divjak, D., Levshina, N. & Klavan, J. Cognitive Linguistics: Looking back, looking forward. *Cognitive Linguistics* **27,** 447–463 (2016).
25. Gil, D. Roon ve, DO/GIVE coexpression, and language contact in Northwest New Guinea. *Nusa: linguistic studies of Indonesian and other languages in Indonesia* **62,** 43–100 (2017).
26. Georgakopoulos, T. & Polis, S. The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass* **12.** e12270 LNCO-0727.R1, e12270–n/a (2018).
27. San Roque, L., Kendrick, K. H., Norcliffe, E. & Majid, A. Universal meaning extensions of perception verbs are grounded in interaction. *Cognitive Linguistics* **29,** 371–406 (2018).
28. List, J.-M. *et al.* CLICS²: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology* **22,** 277–306 (2018).
29. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3** (2016).
30. List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T. & Forkel, R. Sequence comparison in computational historical linguistics. *Journal of Language Evolution* **3** (2018).
31. Merkel, D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.* **2014.** ISSN: 1075-3583 (Mar. 2014).
32. List, J.-M., Cysouw, M. & Forkel, R. *Concepticon. A resource for the linking of concept lists* in *Proceedings of the Tenth International Conference on Language Resources and Evaluation* LREC 2016Portorož, May 23–28, 2016 (eds Chair), N. C. (*et al.*) (European Language Resources Association (ELRA), 2016), 2393–2400.
33. Matisoff, J. A. *The Sino-Tibetan Etymological Dictionary and Thesaurus project* https://stedt.berkeley.edu/ (University of California, Berkeley, 2015).
34. *Zàngmiǎnyǔ yǔyīn hé cíhuì* (ed Sūn, H. 孙.) (Zhōngguó Shèhuì Kēxué, 1991).

35. Solc, T. & Burke, S. M. *Unidecode – ASCII transliterations of Unicode text* 2019.
36. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Physics reports* **659,** 1–44 (2016).
37. Emmons, S., Kobourov, S., Gallant, M. & Börner, K. Analysis of network clustering algorithms and cluster quality metrics at scale. *PloS one* **11,** e0159161 (2016).
38. Holland, P. W. & Leinhardt, S. Transitivity in structural models of small groups. *Comparative group studies* **2,** 107–124 (1971).
39. Newman, M. E. J. *Networks. An Introduction* (Oxford University Press, Oxford, 2010).
40. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 1118–1123 (2008).
41. Forkel, R. & Bank, S. *CLLD: A toolkit for cross-linguistic databases* Jena, 2018. https://doi.org/10.5281/zenodo.1186271.
42. Bodt, T. A. & List, J.-M. Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in Western Kho-Bwa langauges. *Papers in Historical Phonology* **4,** 22–44. http://journals.ed.ac.uk/pihph/article/view/3037 (2019).
43. List, J.-M., Mayer, T., Terhalle, A. & Urban, M. *CLICS: Database of Cross-Linguistic Colexifications* Marburg, 2014. http://clics.lingpy.org.
44. *Loanwords in the world's languages* (eds Haspelmath, M. & Tadmor, U.) (de Gruyter, Berlin and New York, 2009).
45. Abraham, B., Sako, K., Kinny, E. & Zeliang, I. in (ed Anonymous) (India, 2005). https://en.wiktionary.org/wiki/Appendix:Tshangla_comparative_vocabulary_list (2019).
46. Allen, B. *Bai Dialect Survey* http://www.sil.org/silesr/2007/silesr2007-012.pdf (SIL International, 2007).
47. Greenhill, S. J. & Gray, R. *Bantu Basic Vocabulary Database* 2015.
48. *Hànyǔ fāngyán cíhuì* 汉语方言词汇 (ed Běijīng Dàxué, 北.) (Wénzì Gǎigé, 1964).
49. Bowern, C. & Atkinson, Q. D. Computational phylogenetics of the internal structure of Pama-Nguyan. *Language* **88,** 817–845 (2012).
50. Castro, A. & Xingwen, P. *Sui Dialect Research* ISBN: 1934-2470 (SIL International, Guiyang, 2015).
51. Chén, Qíguāng 陳其光. *Miáoyáo yǔwén* 苗瑤语文 *[Miao and Yao language]* (Zhōngyāng Mínzú Dàxué 中央民族大学 [China Minzu University Press], Běijīng, 2012).

52. Carling, G. *et al.* Diachronic Atlas of Comparative Linguistics (DiACL). A database for ancient language typology. *PLOS ONE,* 1–20. https://doi.org/10.1371/journal.pone.0205313 (2018).

53. Hale, A. *Clause, sentence, and discourse patterns in selected languages of Nepal. Part IV. Wordlists* (Summer Institute of Linguistics and Tribhuvan University Press, Kathmandu, 1973).

54. Hantgan, A. & List, J.-M. Bangime. Secret language, language isolate, or language island? *Journal of Language Contact.* forthcoming).

55. Huber, R. Q. & Reed, R. B. *Vocabulario comparativo: palabras selectas de lenguas indígenas de Colombia [Comparative vocabulary. Selected words from the indigeneous languages of Columbia]* (Asociatión Instituto Lingüístico de Verano, Santafé de Bogotá, 1992).

56. Key, M. R. & Comrie, B. *The intercontinental dictionary series* (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2016).

57. Kraft, C. H. *Chadic wordlists* (Dietrich Reimer, Berlin, 1981).

58. Kaiping, G. A. & Klamer, M. LexiRumah: An online lexical database of the Lesser Sunda Islands. *PLOS ONE* **13,** 1–29. https://doi.org/10.1371/journal.pone.0205250 (2018).

59. Marrison, G. E. in (ed Matisoff, J.) (University of California, Berkeley, 2015). http://stedt.berkeley.edu/~stedt-cgi/rootcanal.pl/source/GEM-CNL (2018).

60. Mitterhofer, B. *Lessons from a dialect survey of Bena: Analyzing wordlists* (SIL International, 2013).

61. Nagano, Y. & Prins, M. in (University of California, Berkeley, 2013). https://stedt.berkeley.edu/~stedt-cgi/rootcanal.pl/source/YN-RGLD (2019).

62. Dellert, J. & Jäger, G. *NorthEuraLex (Version 0.9)* (Eberhard-Karls University Tübingen, Tübingen, 2017).

63. Robinson, L. C. & Holton, G. Internal classification of the Alor-Pantar language family using computational methods applied to the lexicon. *Language Dynamics and Change* **2,** 123–149 (2012).

64. Satterthwaite-Phillips, D. *Phylogenetic inference of the Tibeto-Burman languages or on the usefuseful of lexicostatistics (and "megalo"-comparison) for the subgrouping of Tibeto-Burman* Stanford, 2011.

65. So-Hartmann, H. Notes on the Southern Chin languages. *LTBA* (1988).

66. Nurse, D. & Phillipson, G. *Tanzania Language Survey* (Department of Foreign Languages and Linguistics, University of Dar es Salaam, Dar es Salaam, 1975).

67. Greenhill, S. J. TransNewGuinea.org: An Misc Database of New Guinea Languages. *PLoS ONE* **10,** e0141563 (2015).

68. Tryon, D. T. & Hackman, B. D. *Solomon islands languages. An internal classification C* **72** (Pacific Linguistics, Canberra, 1983).

69. Z'graggen, J. A. *A comparative word list of the Northern Adelbert Range Languages, Madang Province, Papua New Guinea* (Pacific Linguistics, Canberra, 1980).

70. Hua, X., Greenhill, S. J., Cardillo, M., Schneemann, H. & Bromham, L. The ecological drivers of variation in global language diversity. *Nature Communications* **10,** 1–10 (2019).

71. Rama, T., List, J.-M., Wahle, J. & Jäger, G. *Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?* in *Proceedings of the North American Chapter of the Association of Computational Linguistics* NAACL 18New Orleans, June 1–6, 2018 (2018), 393–400.

72. Himsolt, M. *GML: A portable graph file format* tech. rep. (Universität Passau, 1997). https://www.fim.uni-passau.de/fileadmin/files/lehrstuhl/brandenburg/projekte/gml/gml-technical-report.pdf.

73. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13,** 2498–2504 (2003).

74. Hartuv, E. & Shamir, R. A clustering algorithm based on graph connectivity. *Information processing letters* **76,** 175–181 (2000).

75. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008,** P10008 (2008).

76. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95,** 14863–14868 (1998).

77. Zhu, X. & Ghahramani, Z. *Learning from Labeled and Unlabeled Data with Label Propagation* tech. rep. (Carnegie Mellon University, 2002).

78. Hopcroft, J. & Tarjan, R. Algorithm 447: Efficient Algorithms for Graph Manipulation. *Communications of the ACM* **16,** 372–378. http://doi.acm.org/10.1145/362248.362272 (1973).

Datasets

79. List, J.-M., Cysouw, M., Greenhill, S. & Forkel, R. *Concepticon. A Resource for the linking of concept list* Jena, 2018. http://concepticon.clld.org.

80. Hammarström, H., Forkel, R. & Haspelmath, M. *Glottolog* version 3.3. Leipzig, 2018. http://glottolog.org.

81. Yang, C. *Lalo regional varieties: Phylogeny, dialectometry and sociolinguistics* English. PhD dissertation (La Trobe University, Bundoora, 2011).

82. List, J.-M. & Wu, M.-S. *"Chinese Dialect Vocabularies" dataset* CLDF: beidasinitic, Zenodo: https://doi.org/10.5281/zenodo.3265734 (MPI-SHH/DLCE, Jena, 2019).

83. List, J.-M., Wu, M.-S., Epps, P., Castro, A. & Schweikhard, N. E. *"Sui Dialect Research" dataset* CLDF: castrosui, Zenodo: https://doi.org/10.5281/zenodo.3268340 (MPI-SHH/DLCE, Jena, 2019).

84. Forkel, R. & Rzymski, C. *"Diachronic Atlas of Comparative Linguistics" dataset* CLDF: diacl, Zenodo: https://doi.org/10.5281/zenodo.3268395 (MPI-SHH/DLCE, Jena, 2019).

85. Tresoldi, T., List, J.-M., Yang, C. & Pepper, S. *"Lalo Regional Varieties" dataset* CLDF: yanglalo, Zenodo: https://doi.org/10.5281/zenodo.3268582 (MPI-SHH/DLCE, Jena, 2019).

86. List, J.-M., Bodt, T. A., Wu, M.-S. & Schweikhard, N. E. *"Lexical Cognates in Western Kho-Bwa" dataset* CLDF: bodthkhobwa, Zenodo: https://doi.org/10.5281/zenodo.3267095 (MPI-SHH/DLCE, Jena, 2019).

87. List, J.-M., Tresoldi, T. & Satterthwaite-Phillips, D. *"Phylogenetic Inference of the Tibeto-Burman Languages" dataset* CLDF: satterthwaitetb, Zenodo: https://doi.org/10.5281/zenodo.3266825 (MPI-SHH/DLCE, Jena, 2019).

88. List, J.-M. & Tresoldi, T. *"Tibeto-Burman Phonology and Lexicon" dataset* CLDF: suntb, Zenodo: https://doi.org/10.5281/zenodo.3266836 (MPI-SHH/DLCE, Jena, 2019).

89. Dellert, J., Jäger, G., List, J.-M., Forkel, R. & Tresoldi, T. *"NorthEuraLex" dataset* CLDF: northeuralex, Zenodo: https://doi.org/10.5281/zenodo.3266709 (MPI-SHH/DLCE, Jena, 2019).

90. Johann-Mattis List Mei-Shin Wu, Y. L. *"Sociolinguistic Research in Western Arunachal Pradesh" dataset* CLDF: abrahammonpa, Zenodo: https://doi.org/10.5281/zenodo.3267058 (MPI-SHH/DLCE, Jena, 2019).

91. List, J.-M. *"Bai Dialect Survey" dataset* CLDF: allenbai, Zenodo: https://doi.org/10.5281/zenodo.3265675 (MPI-SHH/DLCE, Jena, 2019).

92. Greenhill, S. J. *"Bantu Basic Vocabulary Database" dataset* CLDF: bantubvd, Zenodo: https://doi.org/10.5281/zenodo.3265696 (MPI-SHH/DLCE, Jena, 2019).

93. Tresoldi, T. & Bowern, C. *"Computational phylogenetics of the internal structure of Pama-Nguyan" dataset* CLDF: bowernpny, Zenodo: https://doi.org/10.5281/zenodo.2634734 (MPI-SHH/DLCE, Jena, 2019).

94. List, J.-M., Wu, M.-S. & Cooper, D. *"Miao and Yao Language" dataset* CLDF: chenhmongmien, Zenodo: https://doi.org/10.5281/zenodo.3268372 (MPI-SHH/DLCE, Jena, 2019).

95. Rzymski, C., List, J.-M. & Morozova, N. *"Word Lists of Languages in Nepal" dataset* CLDF: halenepal, Zenodo: https://doi.org/10.5281/zenodo.3268446 (MPI-SHH/DLCE, Jena, 2019).

96. List, J.-M. & Hantgan, A. *"Bangime and Friends" dataset* CLDF: hantganbangime, Zenodo: https://doi.org/10.5281/zenodo.3268497 (MPI-SHH/DLCE, Jena, 2019).

97. List, J.-M., Prokić, J., Cysouw, M. & Bouda, P. *"Comparative Vocabulary" dataset* CLDF: hubercolumbian, Zenodo: https://doi.org/10.5281/zenodo.3265977 (MPI-SHH/DLCE, Jena, 2019).

98. Forkel, R. *"Intercontinental Dictionary Series" dataset* CLDF: ids, Zenodo: https://doi.org/10.5281/zenodo.3265822 (MPI-SHH/DLCE, Jena, 2019).

99. List, J.-M., Cysouw, M. & Bouda, P. *"Chadic Wordlists" dataset* CLDF: kraftchadic, Zenodo: https://doi.org/10.5281/zenodo.2632725 (MPI-SHH/DLCE, Jena, 2019).

100. Kaiping, G. A. & Klamer, M. *"Lexirumah" dataset* CLDF: lexirumah, Zenodo: https://doi.org/10.5281/zenodo.3244244 (MPI-SHH/DLCE, Jena, 2019).

101. List, J.-M. *"Database of Cross-Linguistic Colexifications" dataset* CLDF: logos, Zenodo: https://doi.org/10.5281/zenodo.3264975 (MPI-SHH/DLCE, Jena, 2019).

102. List, J.-M., Wu, M.-S. & Tresoldi, T. *"Naga Languages of North-East India" dataset* CLDF: marrisonnaga, Zenodo: https://doi.org/10.5281/zenodo.3268547 (MPI-SHH/DLCE, Jena, 2019).

103. List, J.-M., Tresoldi, T., Roper, J. & Proos, M. *"Bena Dialect Survey" dataset* CLDF: mitterhoferbena, Zenodo: https://doi.org/10.5281/zenodo.3268556 (MPI-SHH/DLCE, Jena, 2019).

104. Tresoldi, T. *"rGyalrongic Languages Database" dataset* CLDF: naganorgyalrongic, Zenodo: https://doi.org/10.5281/zenodo.3268562 (MPI-SHH/DLCE, Jena, 2019).

105. Tresoldi, T. *"Internal Classification of the Alor-Pantor Language Family" dataset* CLDF: robinsonap, Zenodo: https://doi.org/10.5281/zenodo.3266720 (MPI-SHH/DLCE, Jena, 2019).

106. List, J.-M. *"Notes on the Southern Chin Languages" dataset* CLDF: sohartmannchin, Zenodo: https://doi.org/10.5281/zenodo.3268589 (MPI-SHH/DLCE, Jena, 2019).

107. Tresoldi, T. *"Tanzania Language Survey" dataset* CLDF: tls, Zenodo: https://doi.org/10.5281/zenodo.3266982 (MPI-SHH/DLCE, Jena, 2019).

108. Greenhill, S. & Tresoldi, T. *"TransNewGuinea.org" dataset* CLDF: transnewguineaorg, Zenodo: https://doi.org/10.5281/zenodo.3268577 (MPI-SHH/DLCE, Jena, 2019).

109. Greenhill, S. *"Solomon Islands Languages" dataset* CLDF: tryonsolomon, Zenodo: https://doi.org/10.5281/zenodo.3267005 (MPI-SHH/DLCE, Jena, 2019).

110. Forkel, R. *"The World Loanword Database" dataset* CLDF: wold, Zenodo: https://doi.org/10.5281/zenodo.3267018 (MPI-SHH/DLCE, Jena, 2019).

111. Tresoldi, T. & List, J.-M. *"Madang Comparative Wordlists" dataset* CLDF: zgraggenmadang, Zenodo: https://doi.org/10.5281/zenodo.3267037 (MPI-SHH/DLCE, Jena, 2019).